|  |  |  |
| --- | --- | --- |
| **文章信息** | **摘要** | **动机** |
| ***阅读程度****：*  精读  ***文章标题：***  Adversarial Patch  ***中文标题：***  对抗补丁  ***发表于：***  NIPS2017  ***作者：***  Tom B. Brown, Dandelion Mané | 我们提出了一种方法来在现实世界中创建**通用的、鲁棒的、有针对性的**对抗性图像补丁。  这些补丁是通用的，因为它们可以用来攻击任何场景，健壮是因为它们可以在各种各样的转换下工作，而有针对性是因为它们可以导致分类器输出任何目标类。  这些对抗性补丁可以打印、添加到任何场景、拍照并呈现给图像分类器；即使补丁很小，它们也会导致分类器忽略场景中的其他项目，并报告所选择的目标类。 | 很简单的一个公式，首先要注意作者的理念，我们不是对图片本身进行处理，而是对一个小补丁p进行梯度上升。  A这个代表打补丁方式，p是补丁图片，x是被打补丁的自然样本，l是补丁在自然样本上的位置，t是对补丁进行何种变换（如旋转）。  在训练过程中每次获得一个minibatch x，N张图片，从T分布与L分部中也抽样N中变换与N种位置，这样就获得了N中打补丁方式，在N张图片上实施N中打补丁方式就获得了向神经网络中输入的内容，然后通过对p进行梯度回传最大化损失函数。  注意整个过程中只对一个补丁进行不断的更新优化，期望达到的效果是这个补丁被施加任何变换（鲁棒性），被放在任何图片上的任何位置（通用性），都能诱导神经网络将图片分类为这个给定类别（有针对性）。    还有一点就是，本文建议对抗补丁的初始状态的选择从一个图片开始，如希望将样本攻击为“面包机类”那么需要是一个面包机的图片，在更新过程中要限制p与的无穷范数小于epsilon，这意味着最后得到的补丁p与一个面包机有一定的语义相似性。 |